Mila

Deep Learning for AI

Yoshua Bengio

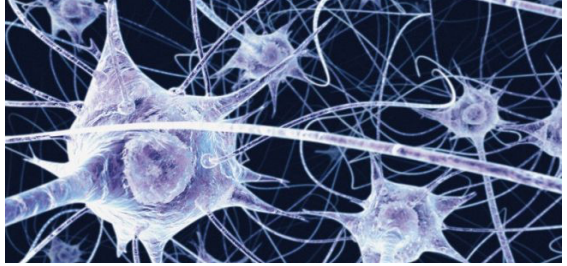
Université de Montréal

TURING AWARD LECTURE
HEIDELBERG LAUREATE FORUM
23 SEPTEMBER 2019

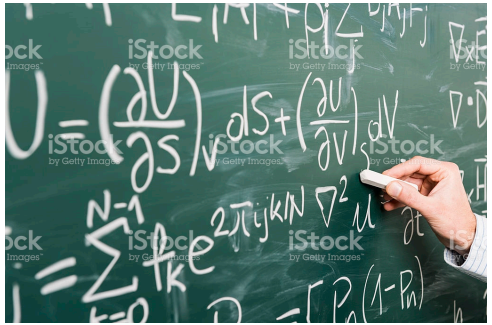
CIFAR
CANADIAN INSTITUTE FOR ADVANCED RESEARCH

ICRA
INSTITUT CANADIEN DE RECHERCHES AVANCÉES

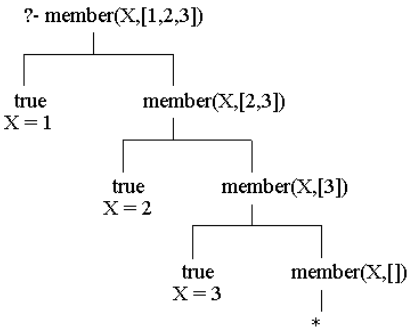
Neural Networks & AI: Underlying Assumption



- There are principles giving rise to intelligence (machine, human or animal) via learning, simple enough that they can be described compactly, similarly to the laws of physics, i.e., our intelligence is not just the result of a huge bag of tricks and pieces of knowledge, but of general mechanisms to acquire knowledge.

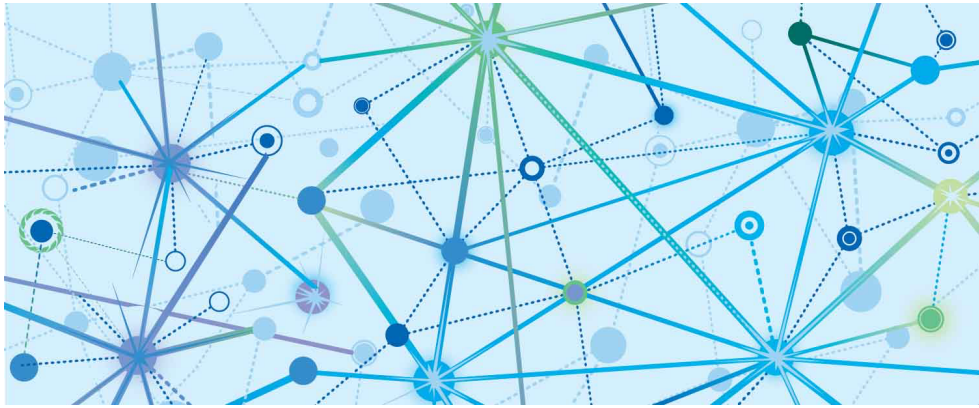


The Machine Learning approach to AI



- **Classical AI, rule-based, symbolic**
  - knowledge is provided by humans
    - but intuitive knowledge (e.g. much of common sense) not communicable
  - machines only do inference
  - no strong learning, adaptation
  - insufficient handling of uncertainty
  - not grounded in low-level perception and action
- **Machine learning tries to fix these problems**
  - succeeded to a great extent
  - higher-level (conscious) cognition still seems out of reach

The Neural Net Approach to AI



- **Brain-inspired**
- Synergy of a large number of simple adaptive computational units
- Focus on **distributed representations**
  - **E.g. word representations** (Bengio et al NIPS'2000)
- View intelligence as arising of combining
  - an objective or reward function
  - an approximate optimizer (learning rule)
  - an initial architecture / parametrization
- End-to-end learning (all the pieces of the puzzle adapt to help each other)

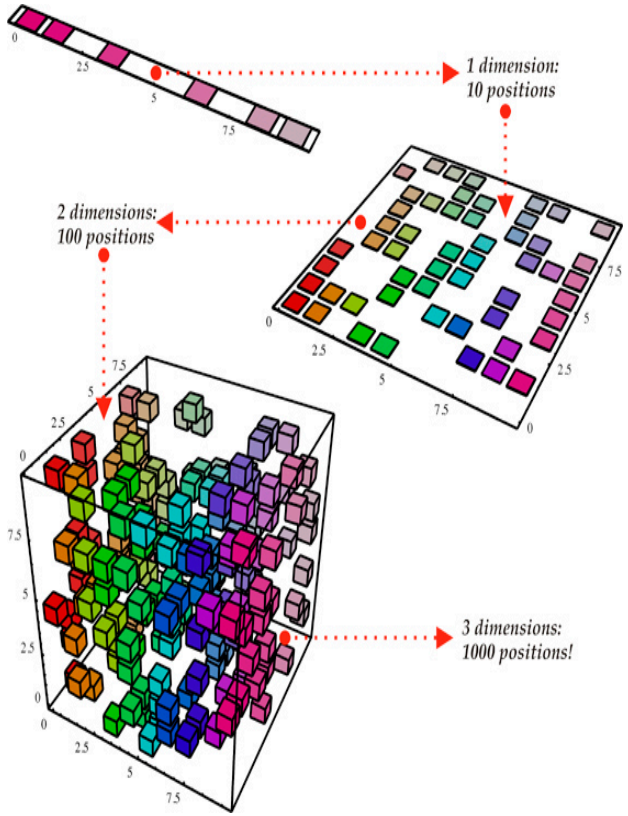
What is Deep Learning about and what did I have to do with it?

- New methods to enable training of deeper networks + theoretical understanding
  - (Bengio et al NIPS'2006): pre-training stacks of auto-encoders before supervised training, greedy supervised and unsupervised pre-training
  - (Glorot & Bengio AISTATS'2010): initialization with near 1 e-values Jacobians
  - (Glorot & Bengio AISTATS'2011): importance of ReLU for training deep nets
- Beyond pattern recognition:
  - Progress in deep unsupervised models, generative models
    - (Vincent et al & Bengio 2008)++: denoising auto-encoders, self-supervised objectives
    - (Goodfellow et al & Bengio 2014): GANs = generative adversarial networks
  - Attention mechanisms, for arbitrary data structures (Bahdanau et al & Bengio 2014)
  - Meta-learning (Bengio & Bengio 1991; many more in last 2 years)

ML 101. What We Are Fighting Against: The Curse of Dimensionality

To generalize
locally, need representative examples for all relevant variations!

Classical solution:
hope for a smooth enough target function, or make it smooth by handcrafting good features / kernel



Bypassing the curse of dimensionality

We need to build compositionality into our ML models

- Just as human languages exploit compositionality to give representations and meanings to complex ideas

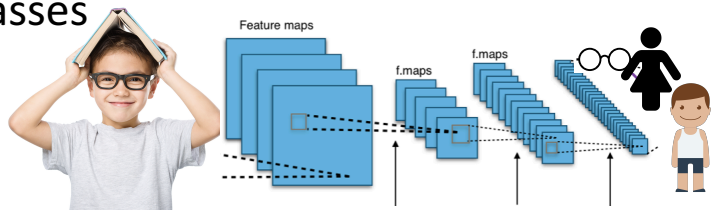
Exploiting compositionality can give an **exponential** gain in representational power

- Distributed representations / embeddings: feature learning
- Deep architecture: multiple levels of feature learning

Prior assumption: compositionality is useful to describe the world around us efficiently

Each feature can be discovered without the need for seeing the exponentially large number of configurations of the other features

- Consider a network whose hidden units discover the following features:
  - Person wears glasses
  - Person is female
  - Person is a child
  - Etc.



- If each of *n* feature requires *O(k)* parameters, need *O(nk)* examples
- Parallel composition of features: can be exponentially advantageous
- Non-parametric methods would require $O(n^d)$ examples

Deep Learning: Learning an Internal Representation

- Unlike other ML methods with either
  - no intermediate representation (linear)
  - or fixed (generally very high-dimensional) intermediate representations (SVMs, kernel machines)

- What is a good representation? Makes other tasks easier.

Exponential advantage of depth

- Expressiveness of deep networks with piecewise linear activation functions: exponential advantage for depth
- *(Montufar et al & Bengio, NIPS 2014)*
- Number of pieces distinguished for a network with depth L and n_i units per layer is at least

$$\left(\prod_{i=1}^{L-1}\left\lfloor\frac{n_i}{n_0}\right\rfloor^{n_0}\right)\sum_{j=0}^{n_0}\binom{n_L}{j}$$

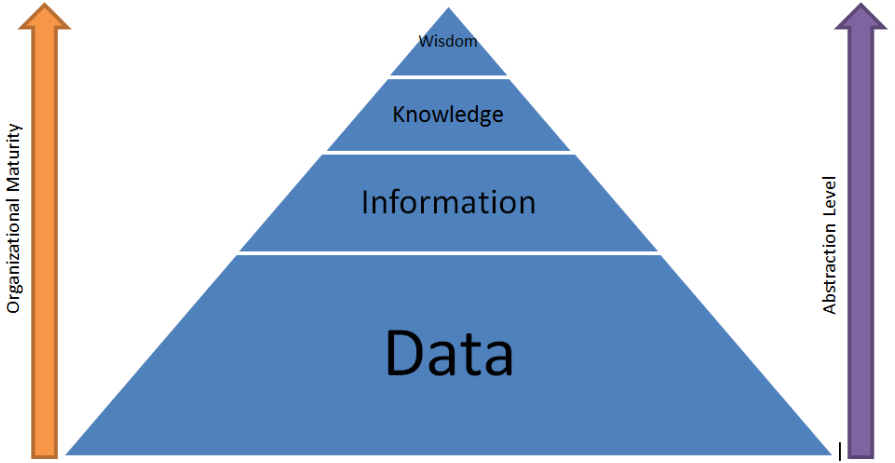
or, if hidden layers have width n and input has size n_0

$$\Omega\left(\left(n/n_0\right)^{(L-1)n_0}\, n^{n_0}\right)$$

Learning Multiple Levels of Abstraction

(Bengio & LeCun 2007)

- The big payoff of deep learning is to allow learning higher levels of abstraction
- Higher-level abstractions **disentangle the factors of variation**, which allows much easier generalization and transfer

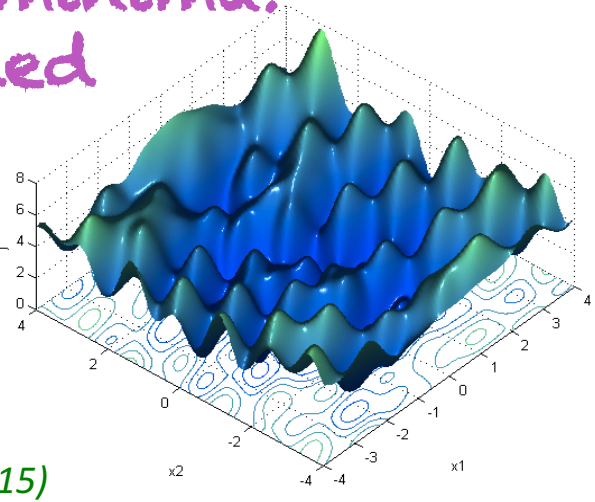


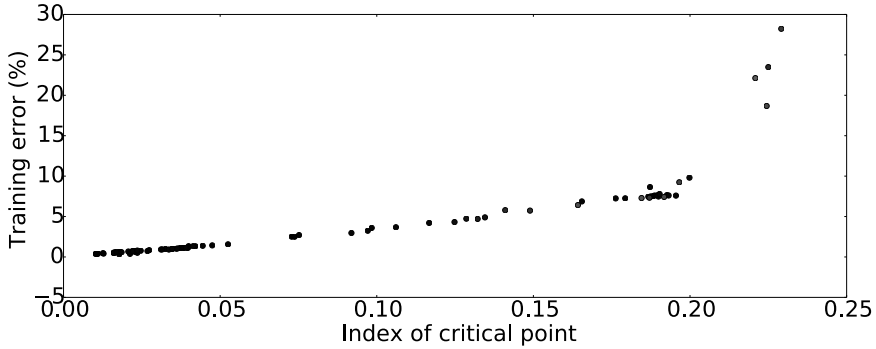
Not so terrible local minima: convexity is not needed

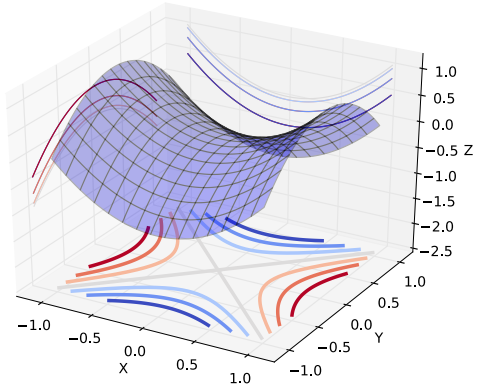
Myth busted:

- Local minima dominate in low-D, but saddle points dominate in high-D
- Most local minima are relatively close to the bottom (global minimum error)

(Dauphin et al NIPS'2014, Choromanska et al AISTATS'2015)





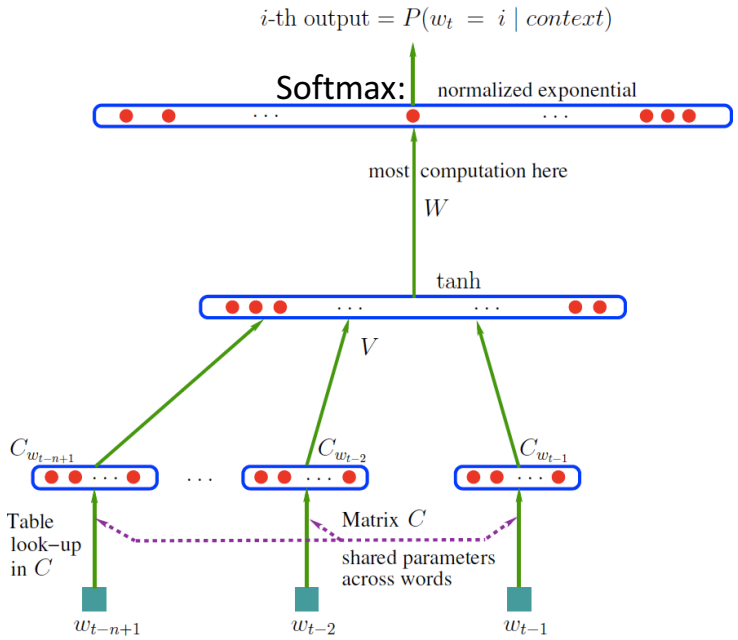


Neural Language Models

- *Bengio et al NIPS'2000 and JMLR 2003 "A Neural Probabilistic Language Model"*



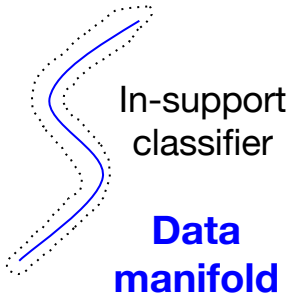
  - Each word represented by a distributed continuous-valued code vector = embedding
  - Generalizes to sequences of words that are semantically similar to training sequences



$$P(w_1, w_2, w_3, \ldots, w_T) = \prod_t P(w_t | w_{t-1}, w_{t-2}, \ldots, w_1)$$

Neural word embeddings - visualization

Classifiers for modeling distributions



- We were inspired by the work of Gutmann & Hyvarinen using probabilistic classifiers to estimate energy functions

  Gutmann & Hyvarinen 2012, Noise-Contrastive Estimation

- In high dimension, more relevant then density is whether you are in-support vs out-of-support

- A classifier of in-support vs out-of-support pays a *constant* price (rather than huge) for not putting support at a training example

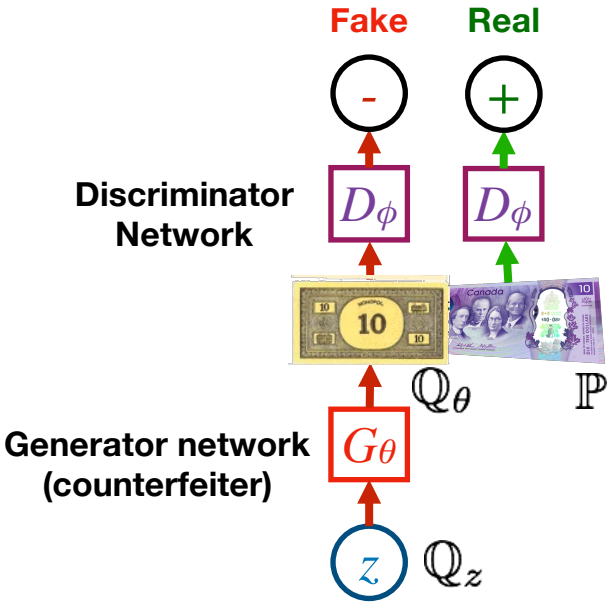
Generative adversarial networks (GANs): a two player game with neural networks

Givens: $\mathbb{Q}_z$ $\mathbb{P}$

Samples from a **target distribution**
(Simple) prior

Player 1: Generator
A neural network with parameters, θ, whose samples **fool the discriminator**

Player 2: Discriminator
Distinguish (classify) real and fake correctly



[Goodfellow et. Al & Bengio, 2014]

Generative Adversarial Networks

Goodfellow et al & Bengio NIPS 2014

2014 2015 2016 2017 2018

this bird is red with white and has a very short beak

Xu et al 2018, AttnGAN

System 1 vs System 2 Cognition

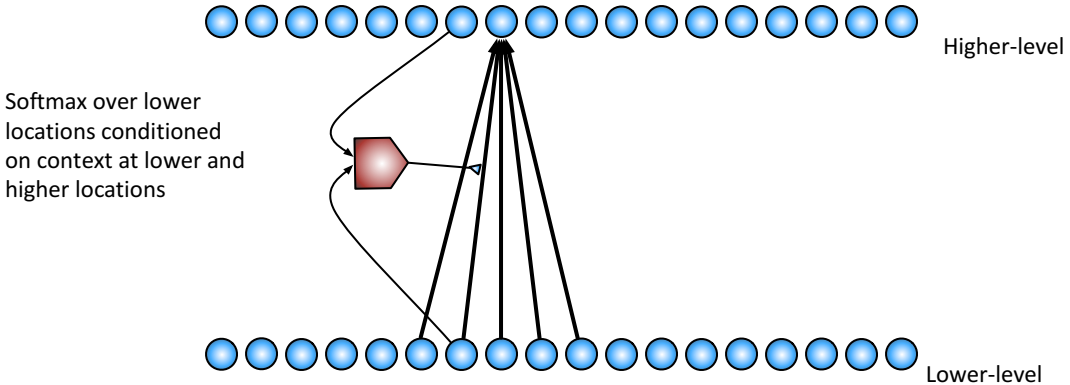
Two systems (and categories of cognitive tasks):

- **System 1**
  - Cortex-like (state controller and representations)
  - intuitive, fast heuristic, UNCONSCIOUS, non-linguistic
  - what current DL does quite well
- **System 2**
  - Hippocampus (memory) + prefrontal cortex
  - slow, logical, sequential, CONSCIOUS, linguistic, algorithmic
  - what classical symbolic AI was trying to do
- **Grounded language learning**: combine both systems

The Attention Revolution in Deep Learning

- **Attention mechanisms exploit GATING units,** have unlocked a breakthrough in machine translation:

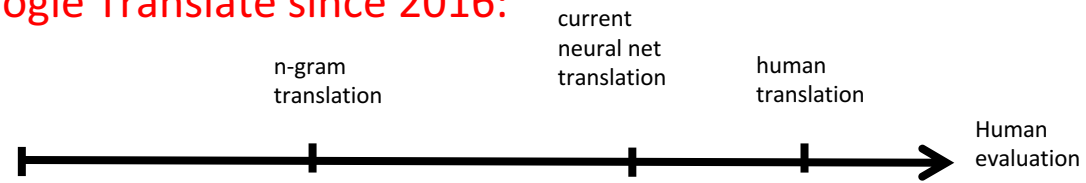
Neural Machine Translation (ICLR'2015)



Attention enables:

- Differentiable memory access
- Operating on sets
- Long-term dependencies
- Self-attention, transformers, SOTA
- Consciousness

- In Google Translate since 2016:

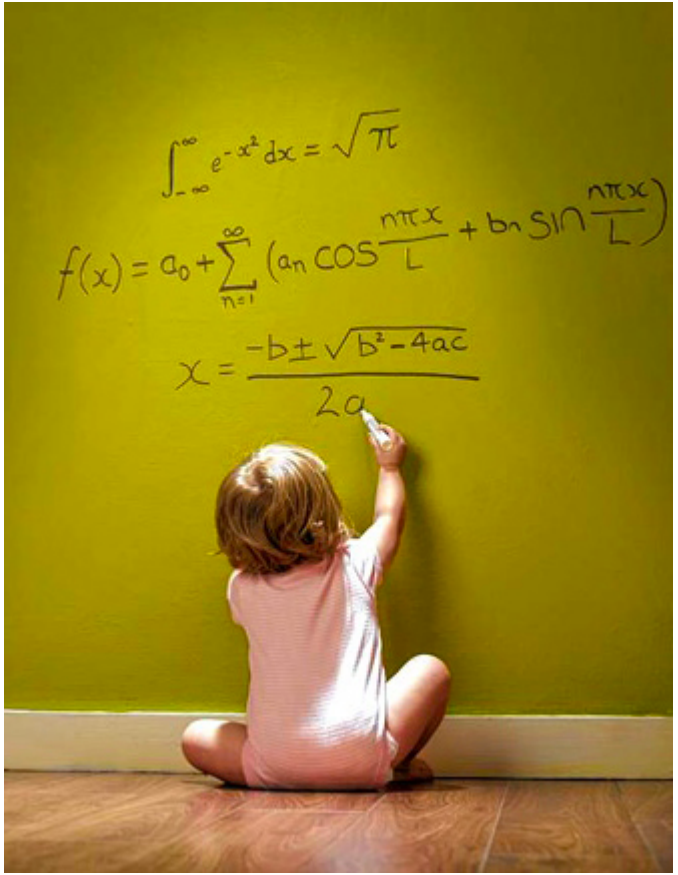


Still Far from Human-Level AI

- Industrial successes mostly based on **supervised** learning requiring lots of human-labeled data implicitly defining the relevant high-level abstractions.
- Learning relatively superficial clues, sometimes not generalizing well outside of training contexts, easy to fool trained networks:

Humans outperform machines at unsupervised learning

- Humans are very good at unsupervised learning, e.g. a 2 year old knows intuitive physics
- Babies construct an approximate but sufficiently reliable model of physics, how do they manage that? Note that they interact with the world, not just observe it.



Learning « How the world ticks »

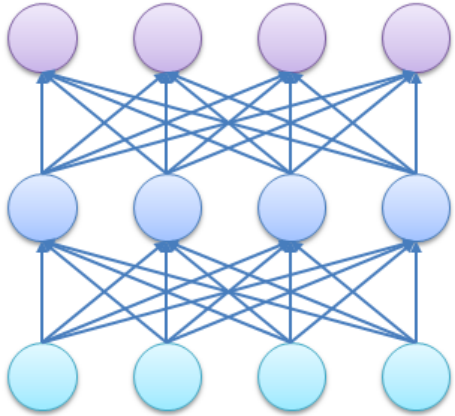
- So long as our machine learning models « cheat » by relying only on superficial statistical regularities, they remain vulnerable to out-of-distribution examples

- Humans generalize better than other animals thanks to a more accurate internal model of the **underlying causal relationships**

- To predict future situations (e.g., the effect of planned actions) far from anything seen before while involving known concepts, an essential component of reasoning, intelligence and science

How to Discover Good Disentangled Representations

- How to discover abstractions?
- What is a good representation? ***(Bengio et al 2013)***
- Need clues (= priors) to help **disentangle** the underlying factors (**not necessarily statistically independent**), such as
  - Spatial & temporal scales
  - Marginal independence
  - Simple dependencies between factors
    - *Consciousness prior*
  - Causal / mechanism independence
    - *Controllable factors*

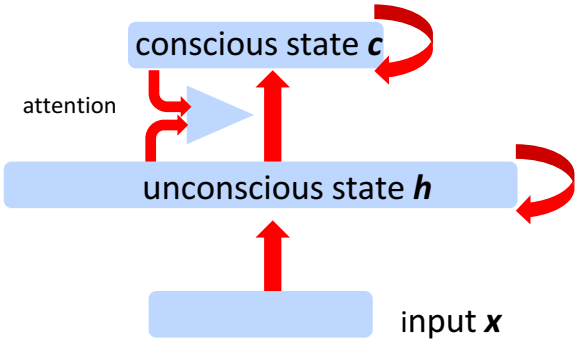


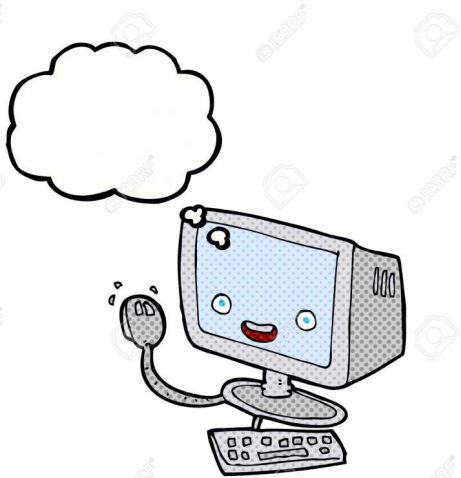
The Consciousness Prior

Bengio 2017, arXiv:1709.08568

- 2 levels of representation:
  - High-dimensional abstract representation space (all known concepts and factors) ***h***

    (not necessarily independent, but with sparse dependencies)
  - Low-dimensional conscious thought ***c***, extracted from ***h***





- ***c*** includes names (keys) and values of factors

Acting to Guide Representation Learning & Disentangling

(E. Bengio et al, 2017; V. Thomas et al, 2017)

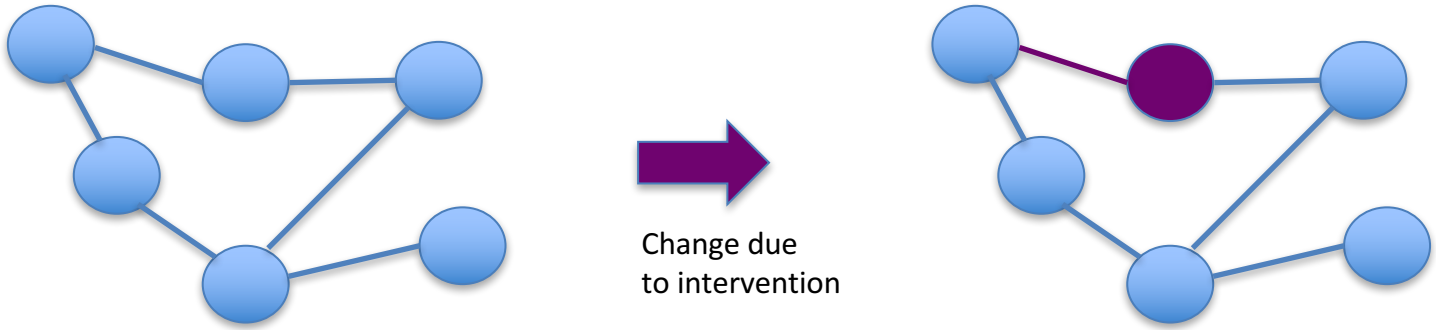
- **Some factors (e.g. objects) correspond to 'independently controllable' aspects of the world**
  - Corresponds to maximizing mutual information between intentions (goal-conditioned policies) and changes in the state (trajectories), conditioned on the current state.
- *Can only be discovered by acting in the world*
  - *Control linked to notion of objects & agents*
  - *Causal but agent-specific & subjective: affordances*

Deep Learning Objective: discover causal representation

- What are the right representations? Causal variables explaining the data
- How to discover them?
- How to discover their causal relationship, the causal graph?

Separating Knowledge in Small Pieces

- Pieces which can be re-used combinatorially
- Pieces which are stable vs nonstationary, subject to interventions



Mila

Missing from Current ML: Understanding & Generalization Beyond the Training Distribution

- Learning theory only deals with generalization within the same distribution
- Models learn but do not generalize well (or have high sample complexity when adapting) to modified distributions, non-stationarities, etc.
- Poor reuse, poor modularization of knowledge

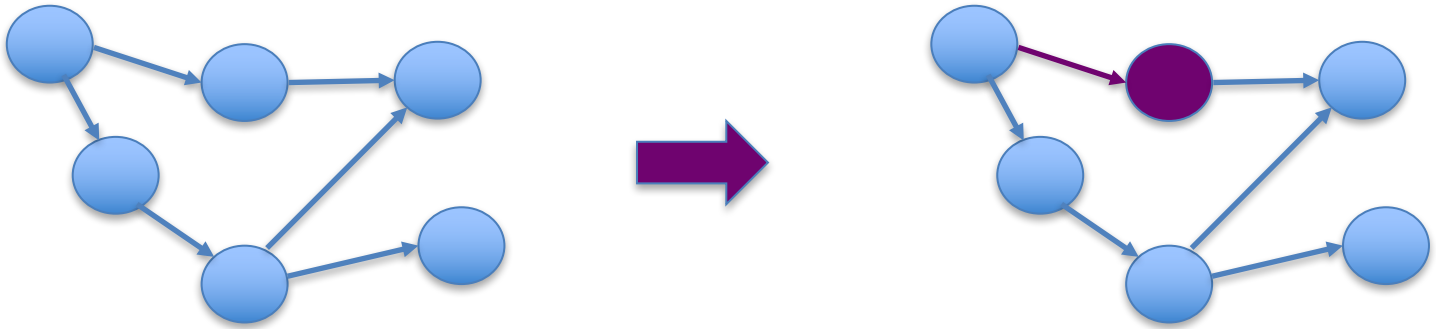
Beyond iid: Hypotheses about how the environment changes
Independent Mechanisms and the Small Change Hypothesis

- Independent mechanisms:
  - changing one mechanism does not change the others *(Peters, Janzig & Scholkopf 2017)*
- Small change:
  - Non-stationarities, changes in distribution, involve few mechanisms (e.g. the result of a single-variable intervention)

Mila

Small Change in the Right Space

Distribution change: only one or a few mechanisms change



Before: eyes open

After: eyes closed,
totally different in pixel space,
small change in object space

Under the right parametrization, few parameters need to change after an intervention

Mila

Turning a Hindrance into a Useful Signal

ArXiv paper, Bengio et al 2019: *A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms*

- Changes in distribution (nonstationarities in agent learning, transfer scenarios, etc) are seen as a bug in ML, a challenge
- Turn them into a feature, an asset, to help discover causal structure, or more generally to help **factorize knowledge:**
- **Tune knowledge factorization (e.g. causal structure) to maximize fast transfer**
- *"Nature does not shuffle environments, we shouldn't"* L. Bottou

Mila

Meta-Learning / Learning to Learn

(Bengio et al 1991)

- Generalize the idea of hyper-parameter optimization
  - Inner loop optimization (normal training), a fn of meta-params

$$\theta_t(\omega) = \text{approxmin}_\theta C(\theta, \omega, \mathcal{D}^t_{train})$$

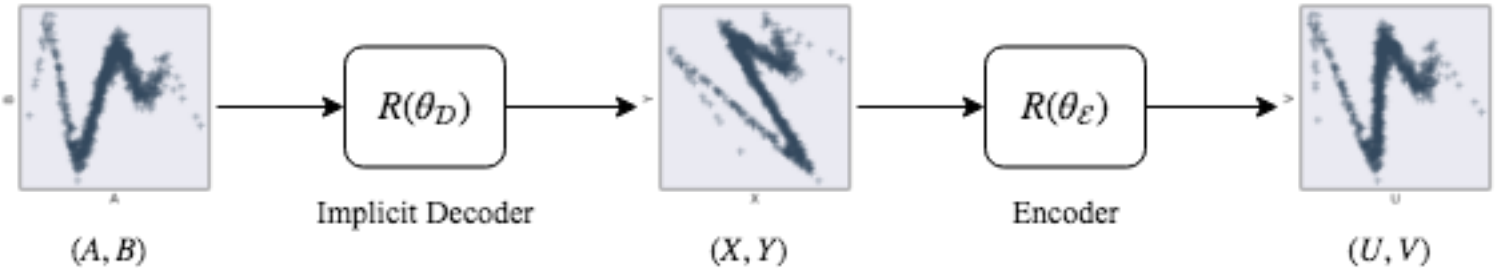
  - Outer loop optimization (meta-training), optimize meta-params

$$\omega = \text{approxmin}_\omega \sum_t L(\theta_t(\omega), \omega, \mathcal{D}^t_{test})$$

- Meta-parameters can be the learning rule itself (Bengio et al 1991; Schmidhuber 1992), learn to optimize
- Meta-learn an objective or reward function, or a shared encoder
- Meta-learning can be used to learn to generalize or transfer
- Can backprop through θ_t, use RL, evolution, or other tricks

Disentangling the Causes

- Realistic settings: causal variables are not directly observed
- Need to learn an encoder which maps raw data to causal space
- Consider both the encoder parameters and the causal graph structural parameters as meta-parameters trained together wrt proposed meta-transfer objective



Simplest possible scenario: linear mixing (rotating decoder) and unmixing (rotating decoder)

Looking Forward

- Build a world model which captures causal effects in abstract space of causal variables, able to quickly adapt to changes in the world and generalize out-of-distribution
- Acting to acquire that knowledge (exploratory behavior)
- Bridging the gap between system 1 and system 2, old neural nets and conscious reasoning, all neural

AI for Social Good

- Beyond developing the next gadget
- AI is powerful, can be misused or bring much good
- Actionable items:
  - Favor ML applications which help the poorest countries, may help with fighting climate change, improve healthcare, education, etc.
  - ***AI Commons***: an organization in construction, which will coordinate, prioritize and channel funding for such applications

XPRIZE · Mila · ITU · HEC PARIS · CHAI Center for Human-Compatible Artificial Intelligence · IEEE Standards

fondation BOTNAR · Pontificia Academia Pro Vita · ocean · Inspired Minds · camera culture · McGovern Foundation

Amir Banifatemi